Linear Regression

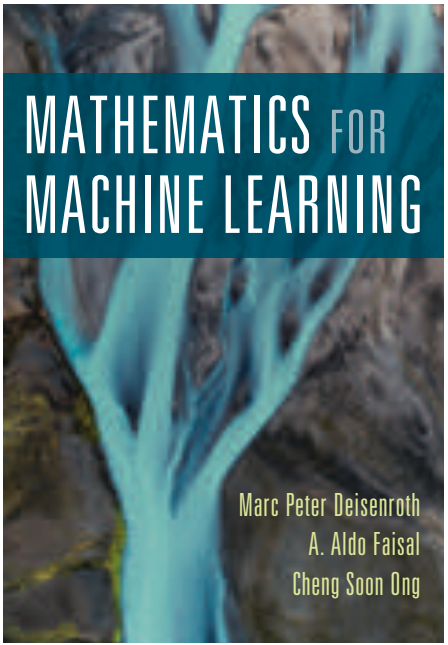
Marc Deisenroth
Centre for Artificial Intelligence
Department of Computer Science
University College London

@mpd37
m.deisenroth@ucl.ac.uk
https://deisenroth.cc

AIMS Rwanda and AIMS Ghana

March/April 2020

Reference



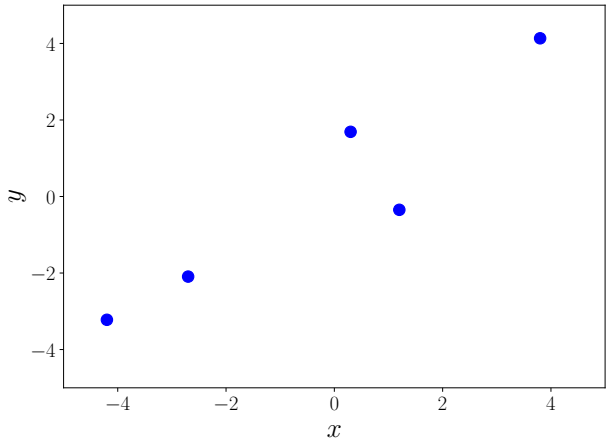
https://mml-book.com

Chapter 9

Regression Problems

Regression (curve fitting)

Given inputs $\boldsymbol{x} \in \mathbb{R}^D$ and corresponding observations $y \in \mathbb{R}$ find a function f that models the relationship between $\boldsymbol{x}$ and y.

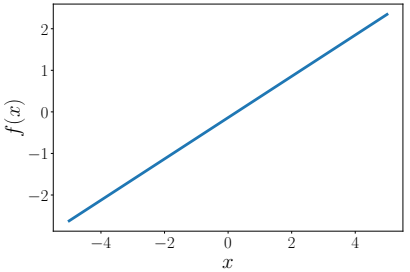


- Typically parametrize the function f with parameters $\boldsymbol{\theta}$
- Linear regression: Consider functions f that are **linear in the parameters**

Linear Regression Functions

- Straight lines

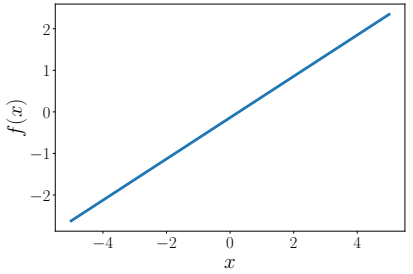
$$y = f(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x = \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix}$$



Linear Regression Functions

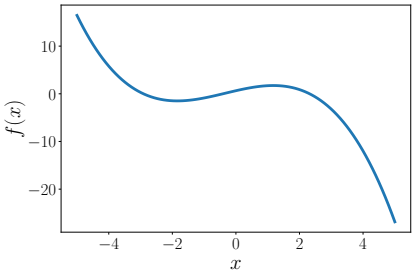
- Straight lines

$$y = f(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x = \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix}$$



- Polynomials

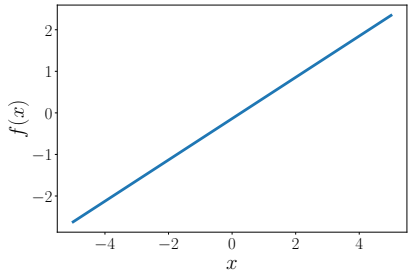
$$y = f(x, \boldsymbol{\theta}) = \sum_{m=0}^{M} \theta_m x^m = \begin{bmatrix} \theta_0 & \cdots & \theta_M \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ x^M \end{bmatrix}$$



Linear Regression Functions

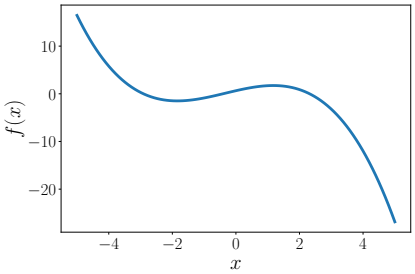
- Straight lines

$$y = f(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x = \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix}$$



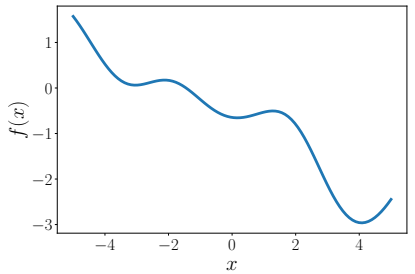
- Polynomials

$$y = f(x, \boldsymbol{\theta}) = \sum_{m=0}^{M} \theta_m x^m = \begin{bmatrix} \theta_0 & \cdots & \theta_M \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ x^M \end{bmatrix}$$



- Radial basis function networks

$$y = f(x, \boldsymbol{\theta}) = \sum_{m=1}^{M} \theta_m \exp\left(-\tfrac{1}{2}(x - \mu_m)^2 \right)$$



Linear Regression Model and Setting

$$y = \boldsymbol{x}^\top \boldsymbol{\theta} + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma^2\right)$$

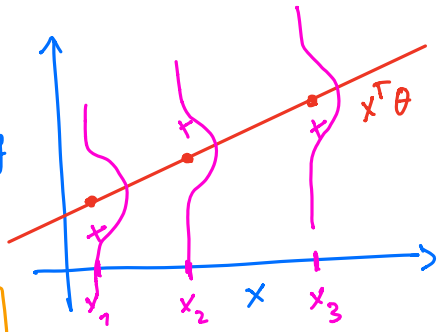
- Given a training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ we seek optimal parameters $\boldsymbol{\theta}^*$

Linear Regression Model and Setting

$$p(y \mid x, \theta) = \mathcal{N}(y \mid x^\top \theta, \sigma^2)$$

$$y = x^\top \theta + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$



- Given a training set $(x_1, y_1), \ldots, (x_N, y_N)$ we seek optimal parameters θ^*
  - **Maximum Likelihood Estimation**
  - **Maximum a Posteriori Estimation**

$p(\epsilon) = \mathcal{N}(0, \sigma^2)$

$a := x^\top \theta$

$y := \epsilon + a$

$\phi(\lambda x + \psi y) = \lambda \phi(x) + \psi \phi(y)$

$\mathrm{var}(\lambda x) = \lambda^2 \mathrm{var}(x)$

$\epsilon + a \Rightarrow p(\epsilon + a) = \mathcal{N}(m, s^2)$

$\mathbb{E}[\epsilon + a] = \underbrace{\mathbb{E}[\epsilon]}_{=0} + a = a$

$\mathrm{var}[\epsilon + a] = \mathrm{var}[\epsilon] = \sigma^2$

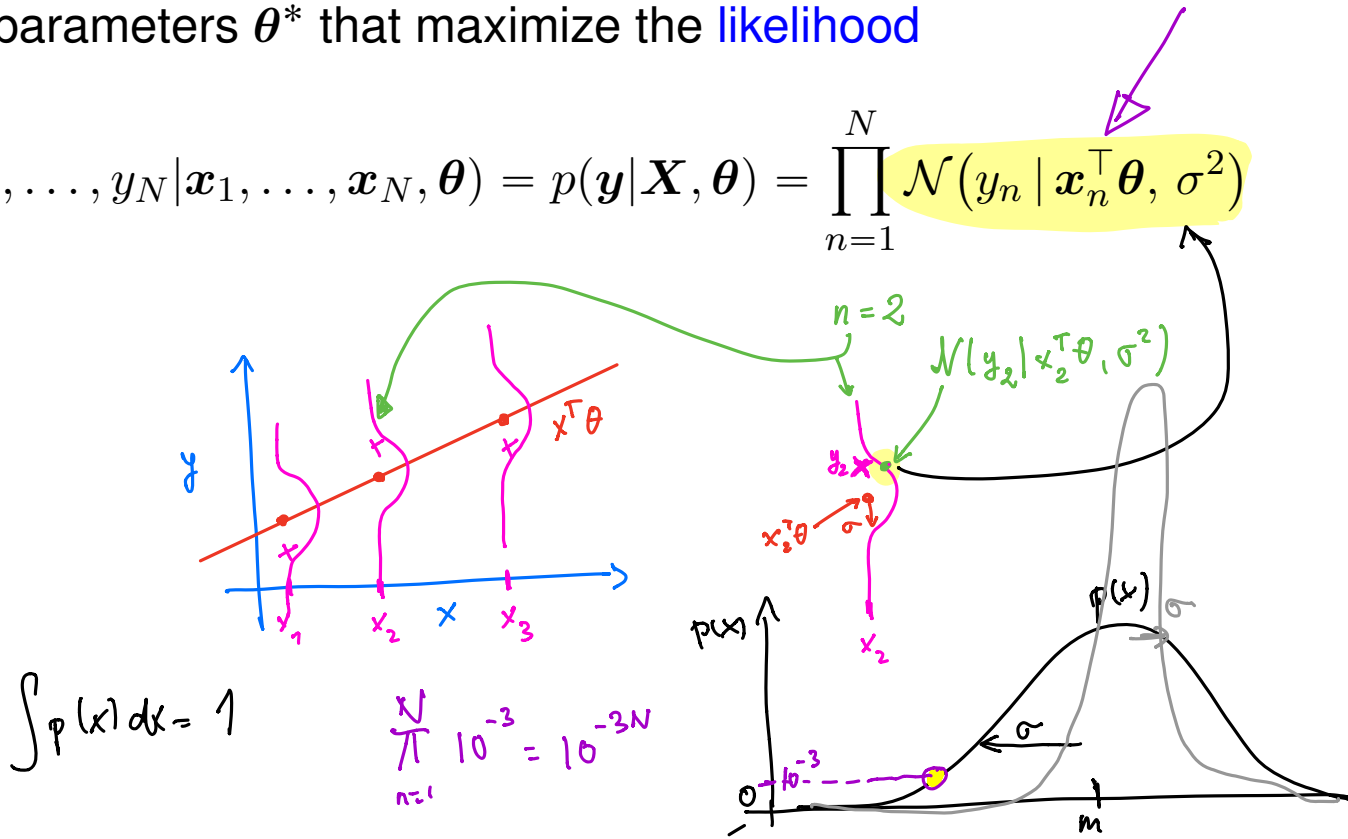
Maximum Likelihood

- Define $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D}$ and $\boldsymbol{y} = [y_1, \ldots, y_N]^\top \in \mathbb{R}^N$
- Find parameters $\boldsymbol{\theta}^*$ that maximize the likelihood

Maximum Likelihood

- Define $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D}$ and $\boldsymbol{y} = [y_1, \ldots, y_N]^\top \in \mathbb{R}^N$
- Find parameters $\boldsymbol{\theta}^*$ that maximize the likelihood

$$p(y_1, \ldots, y_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta}) = p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta}, \sigma^2\right)$$



Maximum Likelihood

- Define $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D}$ and $\boldsymbol{y} = [y_1, \ldots, y_N]^\top \in \mathbb{R}^N$
- Find parameters $\boldsymbol{\theta}^*$ that maximize the likelihood

$$p(y_1, \ldots, y_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta}) = p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right)$$

- Log-transformation ⏩ **Maximize the log likelihood**

Maximum Likelihood

- Define $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D}$ and $\boldsymbol{y} = [y_1, \ldots, y_N]^\top \in \mathbb{R}^N$
- Find parameters $\boldsymbol{\theta}^*$ that maximize the likelihood

$$p(y_1, \ldots, y_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta}) = p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right)$$

- Log-transformation ⏩ **Maximize the log likelihood**

$$\log p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right),$$

$$\log \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right) = -\frac{1}{2\sigma^2}(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 + \text{ const}$$

Maximum Likelihood (2)

With

$$\log \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right) = -\frac{1}{2\sigma^2}(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 + \text{const}$$

we get

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 + \text{const}$$

Maximum Likelihood (2)

With

$$\log \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right) = -\frac{1}{2\sigma^2}(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 + \text{const}$$

we get

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 + \text{const}$$

$$= -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \text{const}$$

Maximum Likelihood (2)

With

$$\log \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right) = -\frac{1}{2\sigma^2}(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 + \text{const}$$

we get

$$\begin{aligned}
\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) &= \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 + \text{const} \\
&= -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \text{const} \\
&= -\frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 + \text{const}
\end{aligned}$$

Maximum Likelihood (2)

With

$$\log \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right) = -\frac{1}{2\sigma^2}(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 + \text{const}$$

we get

$$\begin{aligned}
\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) &= \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \,|\, \boldsymbol{x}_n^\top \boldsymbol{\theta},\, \sigma^2\right) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 + \text{const} \\
&= -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \text{const} \\
&= -\frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 + \text{const}
\end{aligned}$$

- Computing the gradient with respect to $\boldsymbol{\theta}$ and setting it to $\mathbf{0}$ gives the **maximum likelihood estimator** (least-squares estimator)

$$\boldsymbol{\theta}^{\text{ML}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Making Predictions

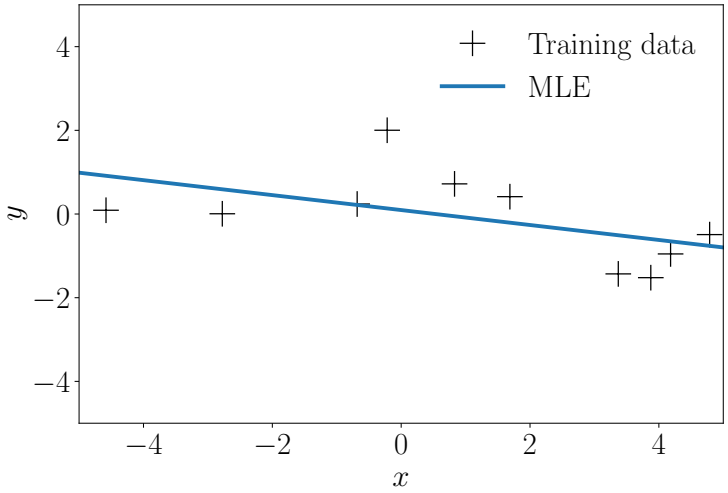
$$y = \boldsymbol{x}^\top \boldsymbol{\theta} + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

Given an arbitrary input $\boldsymbol{x}_*$, we can predict the corresponding observation y_* using the maximum likelihood parameter:

$$p(y_* | \boldsymbol{x}_*, \boldsymbol{\theta}^{\mathsf{ML}}) = \mathcal{N}\left(y_* \,|\, \boldsymbol{x}_*^\top \boldsymbol{\theta}^{\mathsf{ML}}, \sigma^2\right)$$

- Measurement noise variance σ^2 assumed known
- In the absence of noise ($\sigma^2 = 0$), the prediction will be deterministic

Example 1: Linear Functions



$$y = \theta_0 + \theta_1 x + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

- At any query point x_* we obtain the mean prediction as

$$\mathbb{E}[y_*|\boldsymbol{\theta}^{\mathsf{ML}}, x_*] = \theta_0^{\mathsf{ML}} + \theta_1^{\mathsf{ML}} x_*$$

Nonlinear Functions

$$y = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\theta} + \epsilon = \sum_{m=0}^{M} \theta_m x^m + \epsilon$$

- Polynomial regression with features

$$\boldsymbol{\phi}(x) = [1, x, x^2, \ldots, x^M]^\top$$

- Maximum likelihood estimator:

Nonlinear Functions

$$y = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\theta} + \epsilon = \sum_{m=0}^{M} \theta_m x^m + \epsilon$$

- Polynomial regression with features

$$\boldsymbol{\phi}(x) = [1, x, x^2, \ldots, x^M]^\top$$

- Maximum likelihood estimator:

$$\boldsymbol{\theta}^{\mathsf{ML}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}$$

Example 2: Polynomial Regression

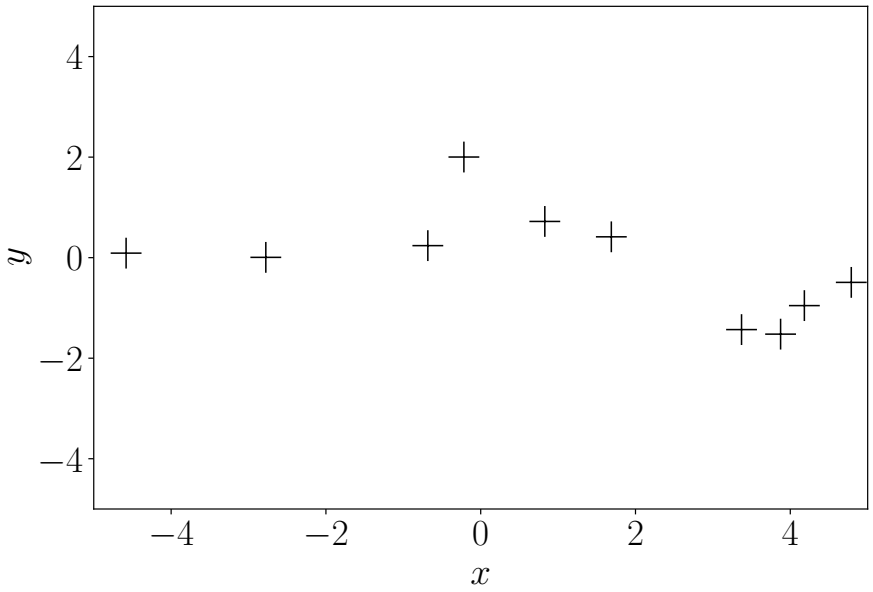


Figure: Training data

Example 2: Polynomial Regression

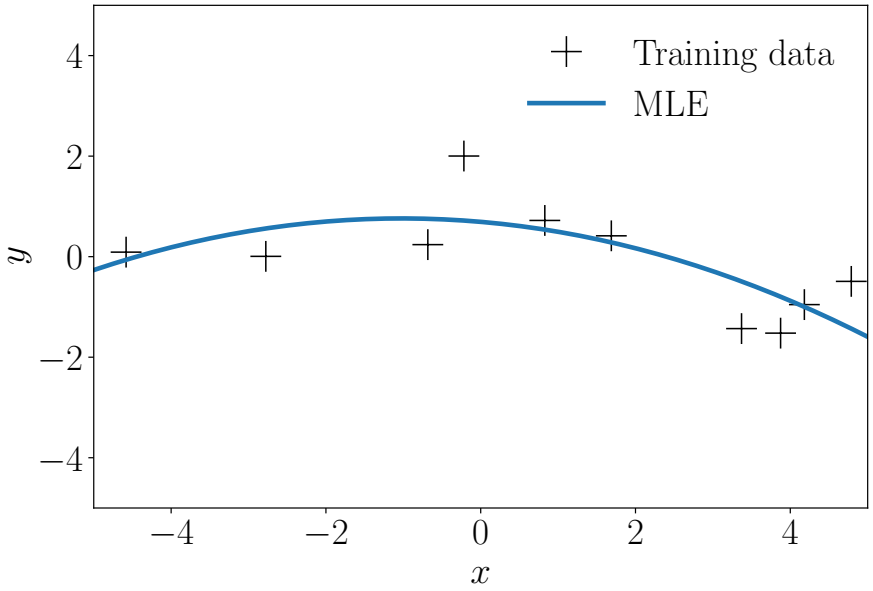


Figure: 2nd-order polynomial

Example 2: Polynomial Regression

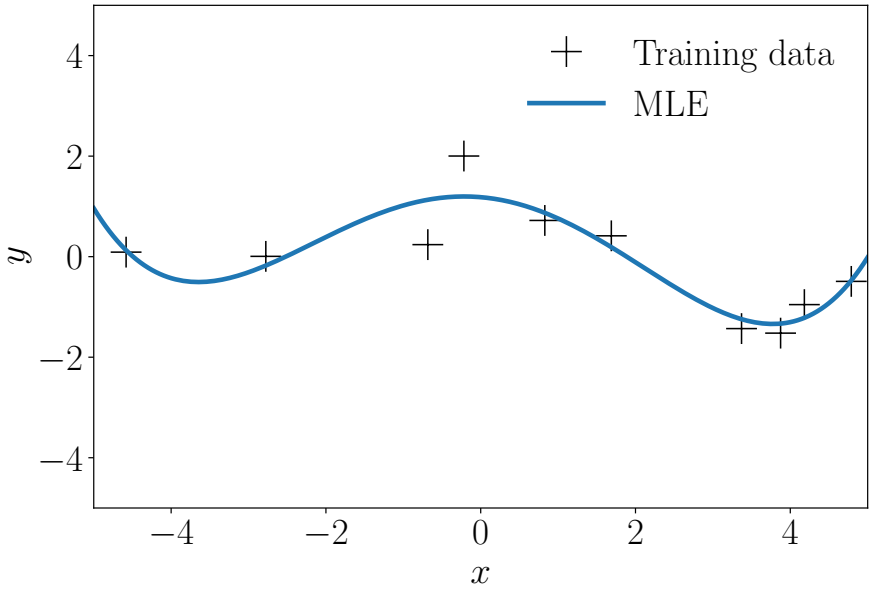


Figure: 4th-order polynomial

Example 2: Polynomial Regression

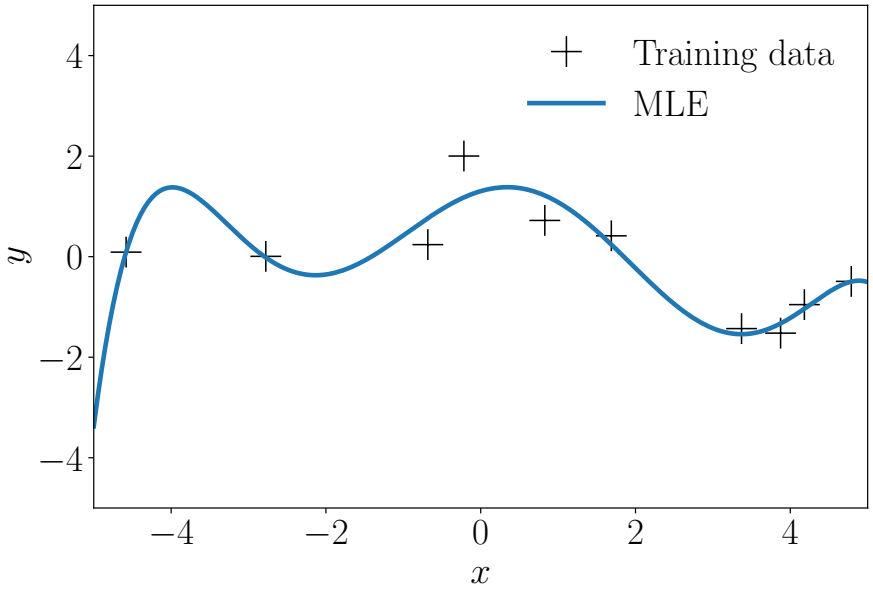


Figure: 6th-order polynomial

Example 2: Polynomial Regression

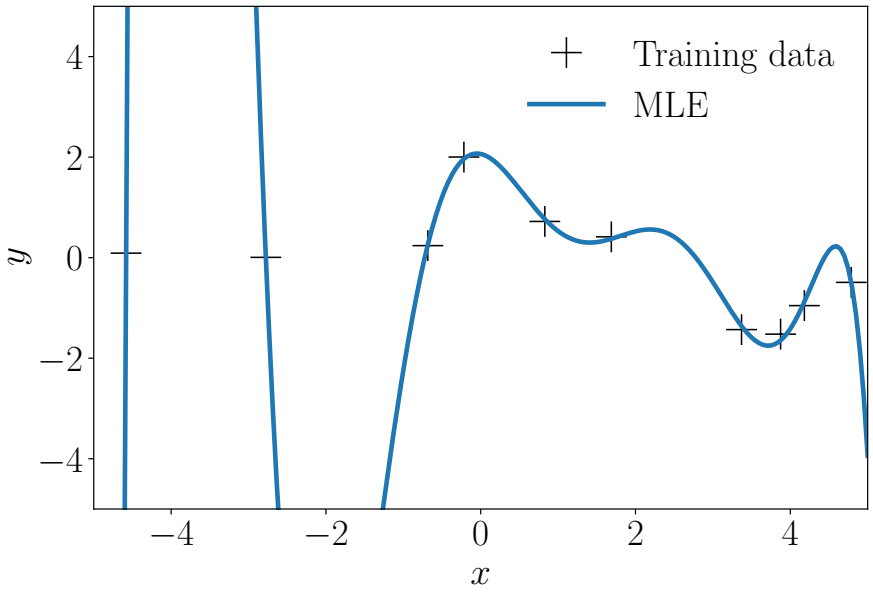


Figure: 8th-order polynomial

Example 2: Polynomial Regression

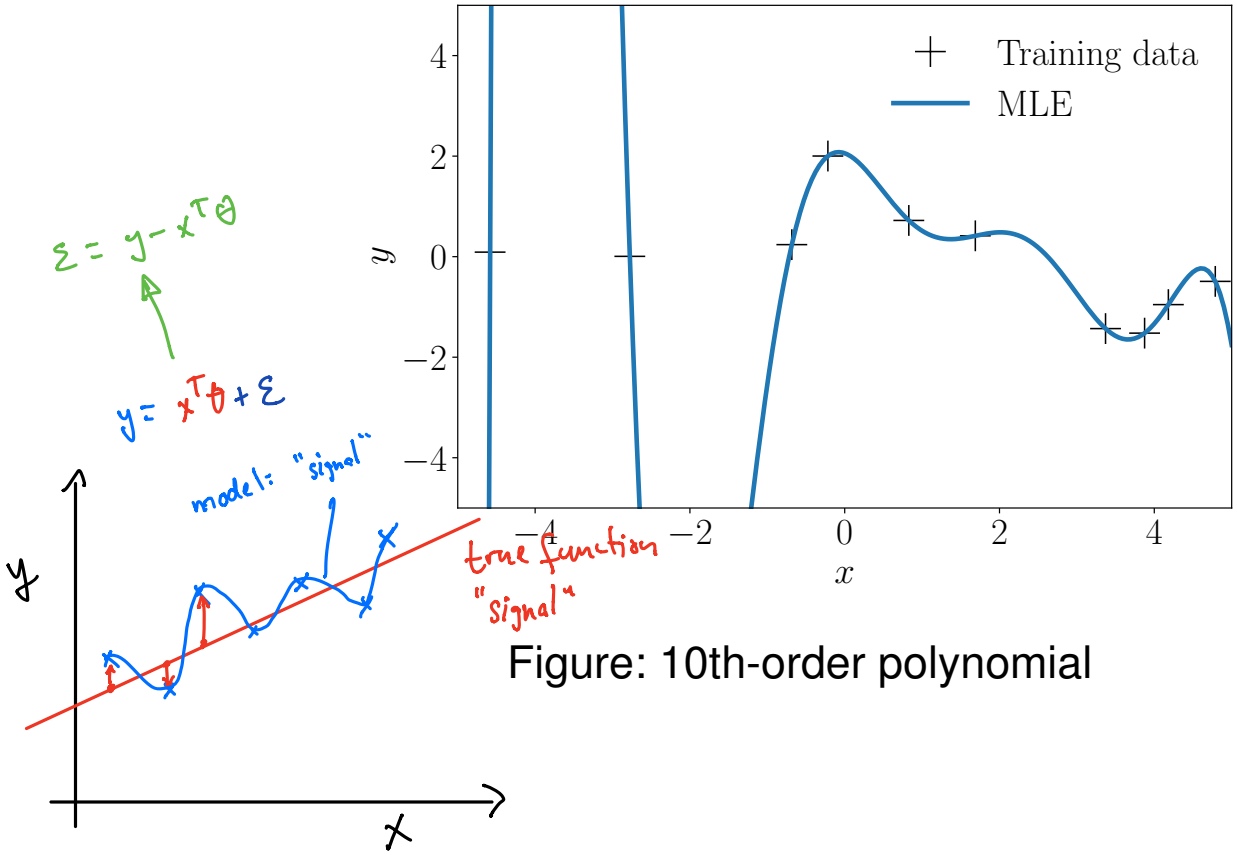
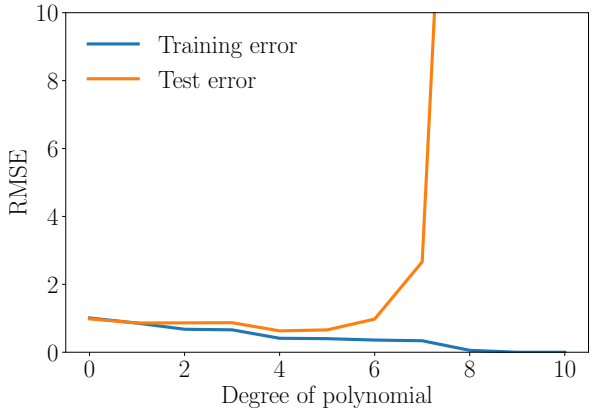


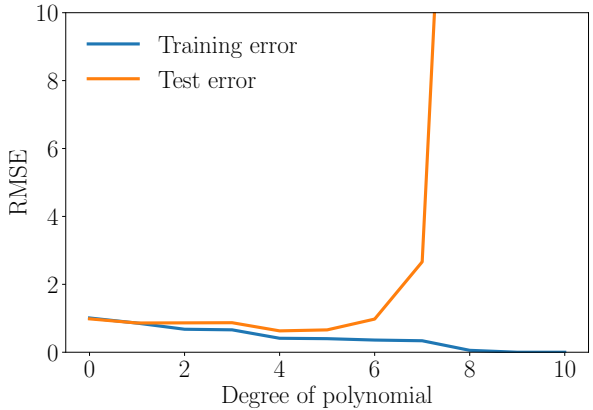
Figure: 10th-order polynomial

Overfitting



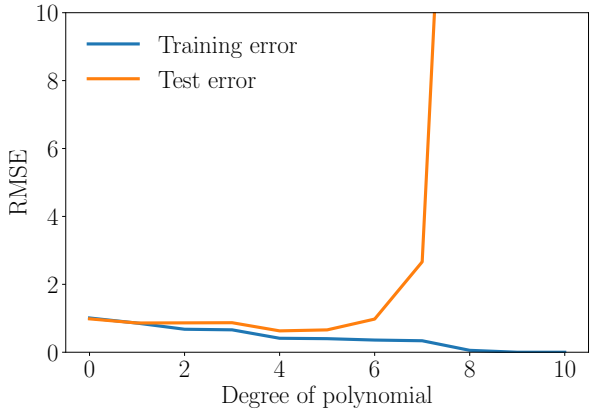
- Training error decreases with higher flexibility of the model

Overfitting



- Training error decreases with higher flexibility of the model
- We are not so much interested in the training error, but in the **generalization error**: How well does the model perform when we predict at previously unseen input locations?

Overfitting



- Training error decreases with higher flexibility of the model
- We are not so much interested in the training error, but in the **generalization error**: How well does the model perform when we predict at previously unseen input locations?
- Maximum likelihood often runs into **overfitting** problems, i.e., we exploit the flexibility of the model to fit to the noise in the data

MAP Estimation

- Empirical observation: Parametric models that overfit tend to have some extreme (large amplitude) parameter values

MAP Estimation

- Empirical observation: Parametric models that overfit tend to have some extreme (large amplitude) parameter values
- Mitigate the effect of overfitting by placing a prior distribution $p(\boldsymbol{\theta})$ on the parameters
  - ⏩ Penalize extreme values that are implausible under that prior

MAP Estimation

- Empirical observation: Parametric models that overfit tend to have some extreme (large amplitude) parameter values
- Mitigate the effect of overfitting by placing a prior distribution $p(\boldsymbol{\theta})$ on the parameters
  - ⏩ Penalize extreme values that are implausible under that prior
- Choose $\boldsymbol{\theta}^*$ as the parameter that maximizes the (log) parameter posterior

$$\log p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \underbrace{\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})}_{\text{log-likelihood}} + \underbrace{\log p(\boldsymbol{\theta})}_{\text{log-prior}} + \text{const}$$

MAP Estimation

- Empirical observation: Parametric models that overfit tend to have some extreme (large amplitude) parameter values
- Mitigate the effect of overfitting by placing a prior distribution $p(\boldsymbol{\theta})$ on the parameters
  - ⏩ Penalize extreme values that are implausible under that prior
- Choose $\boldsymbol{\theta}^*$ as the parameter that maximizes the (log) parameter posterior

$$\log p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \underbrace{\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})}_{\text{log-likelihood}} + \underbrace{\log p(\boldsymbol{\theta})}_{\text{log-prior}} + \text{const}$$

- Log-prior induces a direct penalty on the parameters

MAP Estimation

- Empirical observation: Parametric models that overfit tend to have some extreme (large amplitude) parameter values
- Mitigate the effect of overfitting by placing a prior distribution $p(\boldsymbol{\theta})$ on the parameters
  - ⏩ Penalize extreme values that are implausible under that prior
- Choose $\boldsymbol{\theta}^*$ as the parameter that maximizes the (log) parameter posterior

$$\log p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \underbrace{\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})}_{\text{log-likelihood}} + \underbrace{\log p(\boldsymbol{\theta})}_{\text{log-prior}} + \text{const}$$

- Log-prior induces a direct penalty on the parameters
- **Maximum a posteriori estimate** (regularized least squares)

MAP Estimation (2)

- Gaussian parameter prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \alpha^2 \boldsymbol{I})$
- Log-posterior distribution:

$$\log p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) - \frac{1}{2\alpha^2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \text{const}$$

$$= -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 - \frac{1}{2\alpha^2}\|\boldsymbol{\theta}\|^2 + \text{const}$$

MAP Estimation (2)

- Gaussian parameter prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \alpha^2 \boldsymbol{I})$
- Log-posterior distribution:

$$\log p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) - \frac{1}{2\alpha^2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \text{const}$$

$$= -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 - \frac{1}{2\alpha^2}\|\boldsymbol{\theta}\|^2 + \text{const}$$

- Compute gradient with respect to $\boldsymbol{\theta}$, set it to $\mathbf{0}$

  ⏩ **Maximum a posteriori estimate:**

$$\boldsymbol{\theta}^{\text{MAP}} = (\boldsymbol{X}^\top\boldsymbol{X} + \frac{\sigma^2}{\alpha^2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

Example: Polynomial Regression

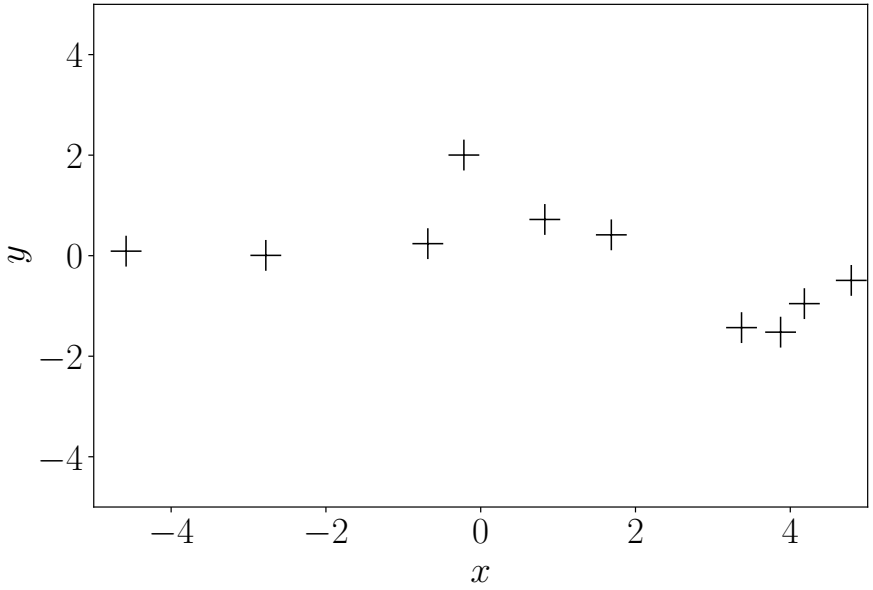


Figure: Training data

Mean prediction:

$$\mathbb{E}[y_* | \boldsymbol{x}_*, \boldsymbol{\theta}^{\text{MAP}}] = \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}^{\text{MAP}}$$

Example: Polynomial Regression

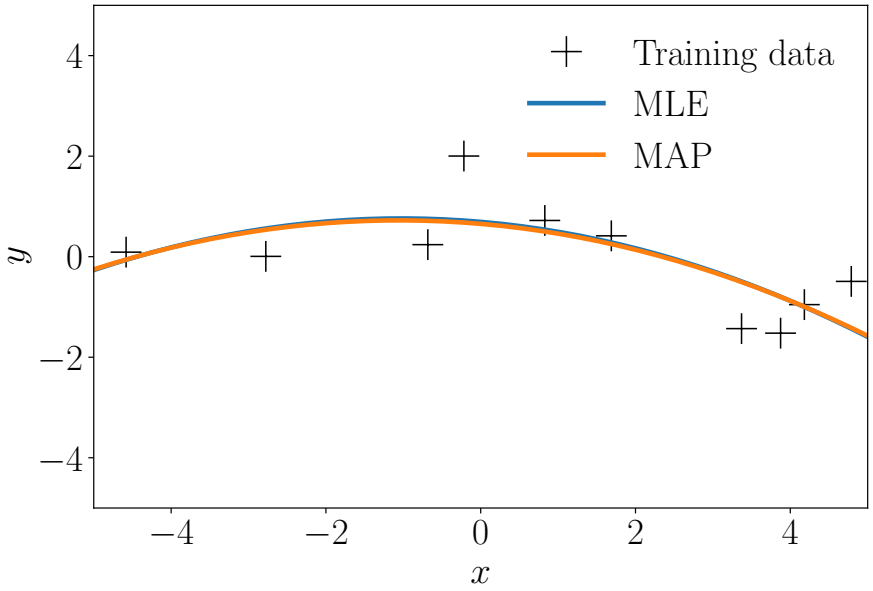


Figure: 2nd-order polynomial

Mean prediction:

$$\mathbb{E}[y_*|\boldsymbol{x}_*, \boldsymbol{\theta}^{\text{MAP}}] = \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}^{\text{MAP}}$$

Example: Polynomial Regression

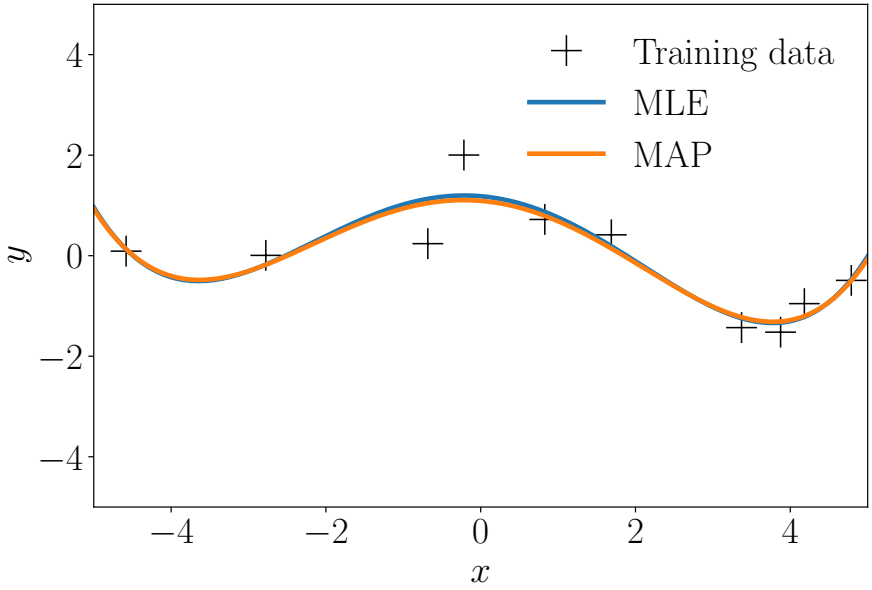


Figure: 4th-order polynomial

Mean prediction:

$$\mathbb{E}[y_*|\boldsymbol{x}_*, \boldsymbol{\theta}^{\text{MAP}}] = \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}^{\text{MAP}}$$

Example: Polynomial Regression

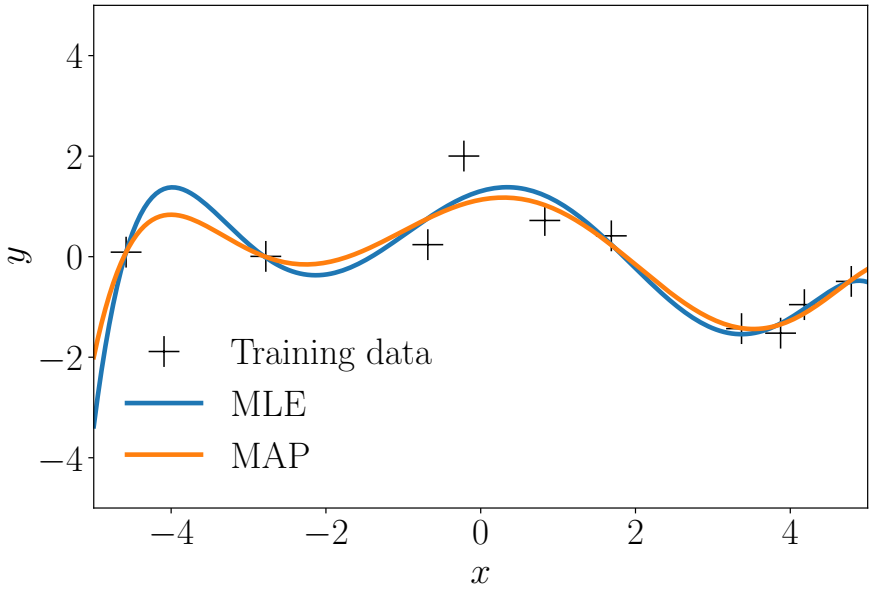


Figure: 6th-order polynomial

Mean prediction:

$$\mathbb{E}[y_*|\boldsymbol{x}_*, \boldsymbol{\theta}^{\text{MAP}}] = \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}^{\text{MAP}}$$

Example: Polynomial Regression

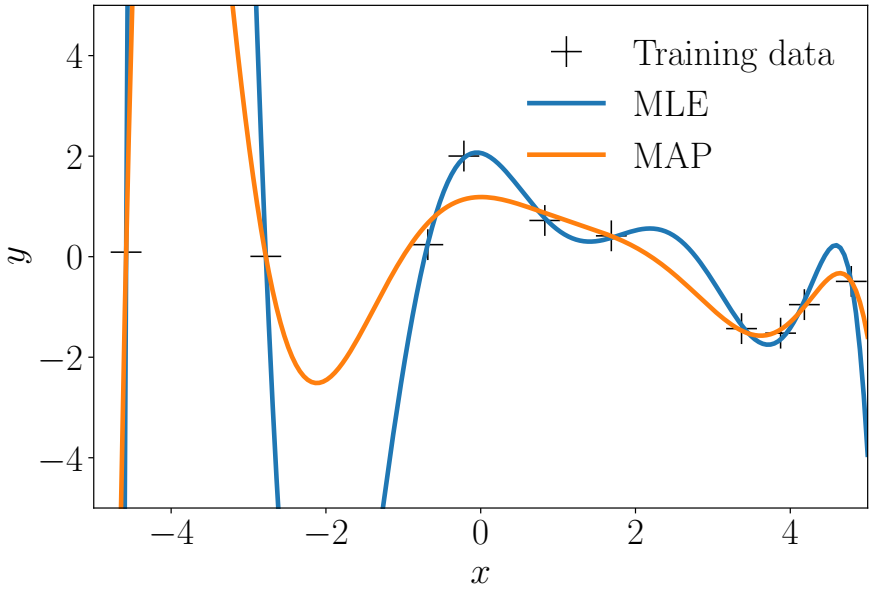


Figure: 8th-order polynomial

Mean prediction:

$$\mathbb{E}[y_*|\boldsymbol{x}_*, \boldsymbol{\theta}^{\text{MAP}}] = \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}^{\text{MAP}}$$

Example: Polynomial Regression

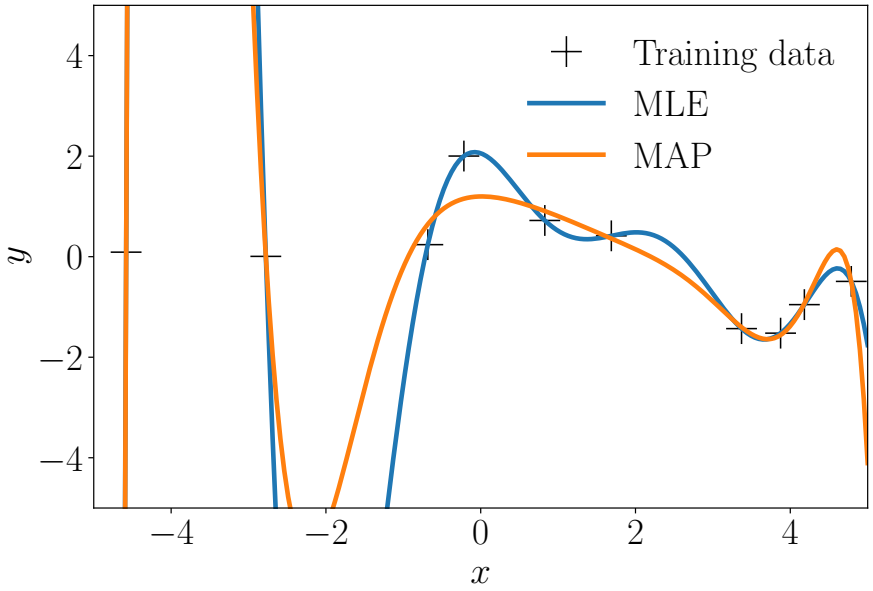
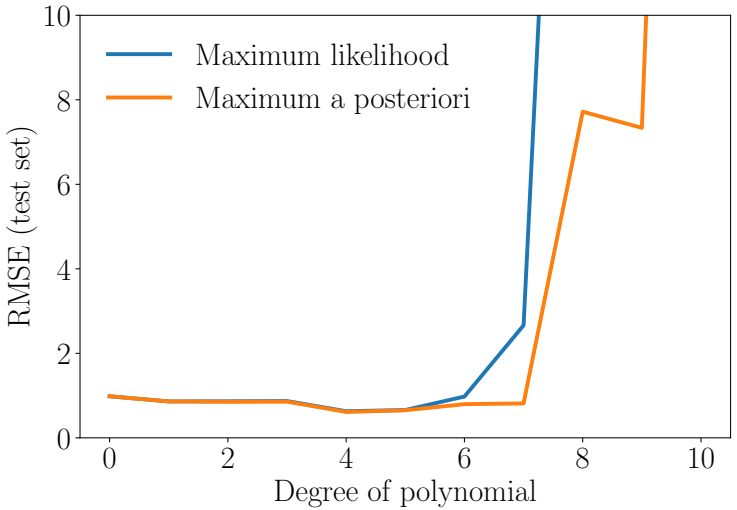


Figure: 10th-order polynomial

Mean prediction:

$$\mathbb{E}[y_*|\boldsymbol{x}_*, \boldsymbol{\theta}^{\text{MAP}}] = \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta}^{\text{MAP}}$$

Generalization Error



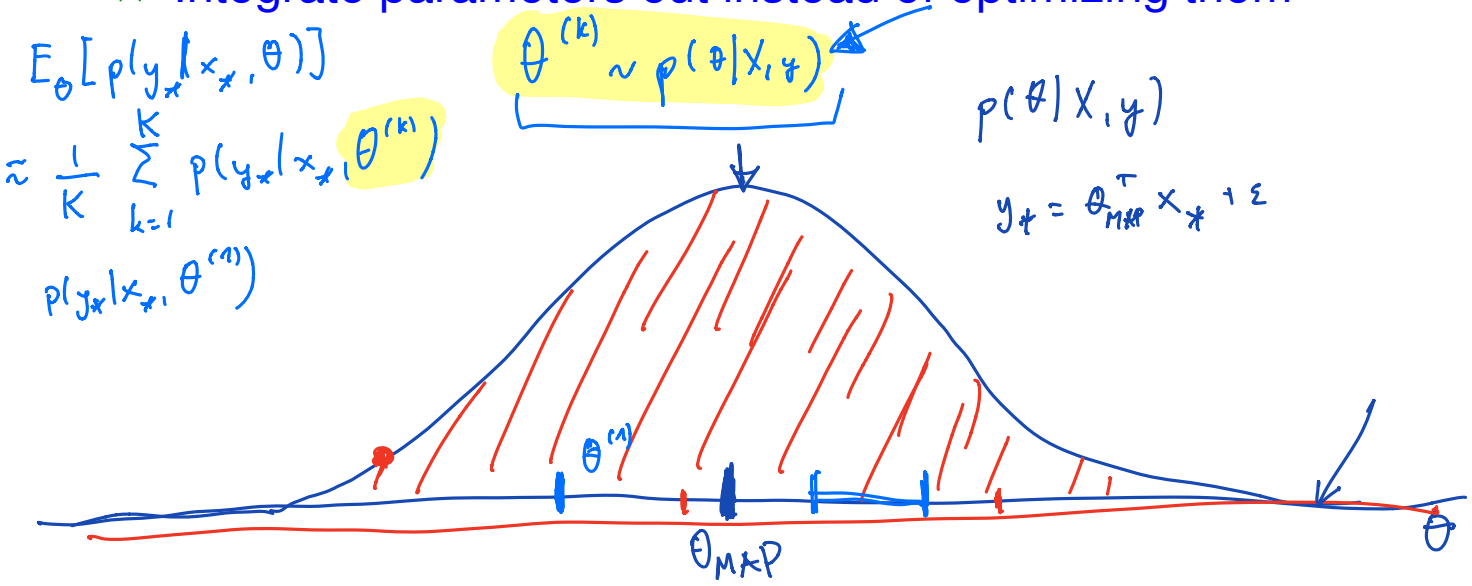
- MAP estimation “delays” the problem of overfitting
- It does not provide a general solution

⏩ Need a more principled solution

Bayesian Linear Regression

$$y = \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta} + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

- Avoid overfitting by not fitting any parameters:
  - Integrate parameters out instead of optimizing them



Bayesian Linear Regression

$$y = \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta} + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

- Avoid overfitting by not fitting any parameters:
  - ⏩ Integrate parameters out instead of optimizing them
- Use a full parameter distribution $p(\boldsymbol{\theta})$ (and not a single point estimate $\boldsymbol{\theta}^*$) when making predictions:

$$p(y_*|\boldsymbol{x}_*) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

  - ⏩ Prediction no longer depends on $\boldsymbol{\theta}$
- Predictive distribution reflects the uncertainty about the "correct" parameter setting

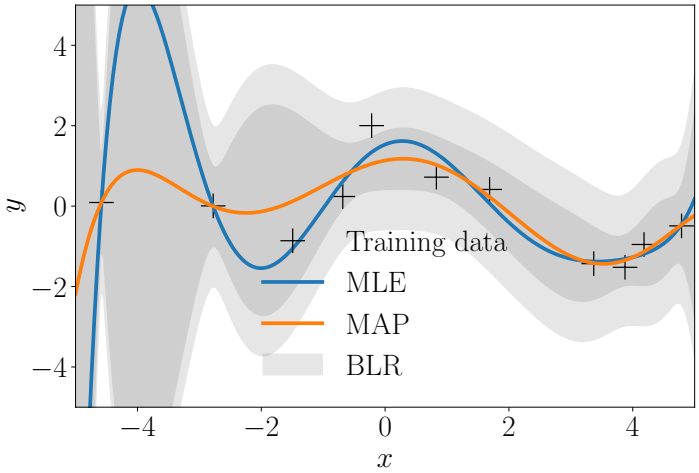
previously: $p(y_*|x_*, \theta^{MAP/ML})$

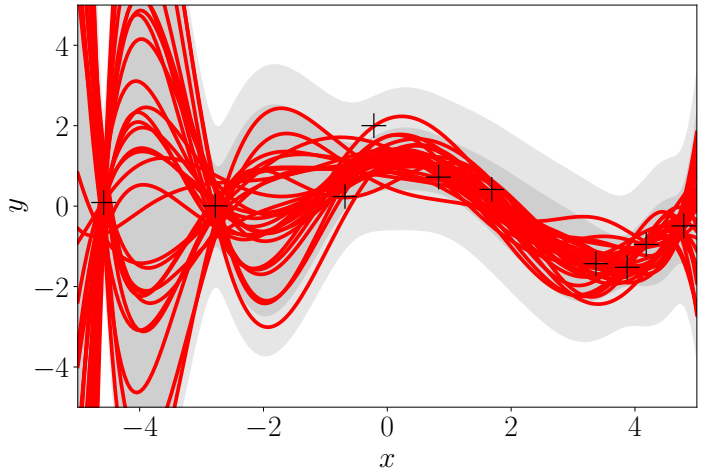
$= \mathbb{E}_\theta[p(y_*|x_*, \theta)]$

plausible values of θ

pred. distribution given a single param. value θ

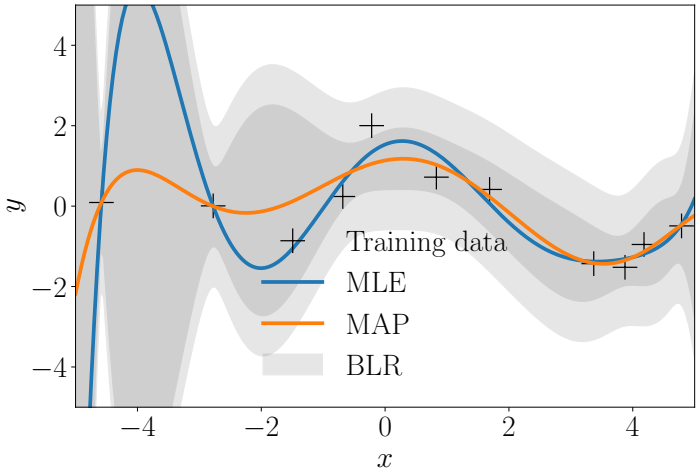
Example

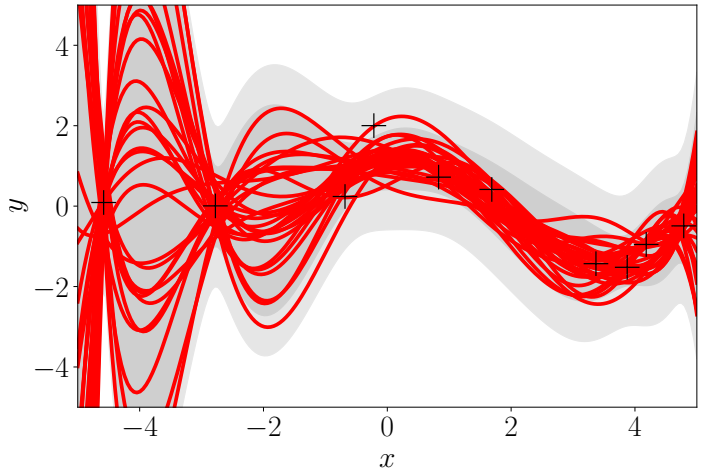




- Light-gray: uncertainty due to noise (same as in MLE/MAP)
- Dark-gray: uncertainty due to parameter uncertainty

Example





- Light-gray: uncertainty due to noise (same as in MLE/MAP)
- Dark-gray: uncertainty due to parameter uncertainty
- Right: Plausible functions under the parameter distribution (every single parameter setting describes one function)

Model for Bayesian Linear Regression

$\exp\left(-\frac{1}{2\sigma^2}(y - x^\top\theta)^2 \quad -\frac{1}{2\alpha^2}\theta^\top\theta\right) \cdot \text{const.}$

$= \ldots \exp\left(-\frac{1}{2}(\theta - m)^\top S^{-1}(\theta - m)\right)$

Prior $\quad p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{S}_0)\,,$

Likelihood $\quad p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y \,|\, \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta}, \sigma^2)$

- Parameter $\boldsymbol{\theta}$ becomes a latent (random) variable
- Prior distribution induces a distribution over plausible functions
- Choose a conjugate Gaussian prior
  - Closed-form computations
  - Gaussian posterior

posterior $\propto$ prior $\times$ likelihood

Given: likelihood

Choose: prior such that the posterior has the same functional form as the prior

e.g.: Gauss. likelihood $\mathcal{N}(y|x^\top\theta, \sigma^2)$

choose: Gauss. prior $\mathcal{N}(\theta|0, \alpha^2 I)$

$\Rightarrow$ posterior: Gaussian $\mathcal{N}(\theta|m, S)$

Parameter Posterior and Predictions

- Prior $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{S}_0)$ is Gaussian ⏩ posterior is Gaussian:
  - ⏩ **Derive this**

$$p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{m}_N, \boldsymbol{S}_N)$$

$$\boldsymbol{S}_N = (\boldsymbol{S}_0^{-1} + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}$$

$$\boldsymbol{m}_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y})$$

$\mathrm{var}_\theta[A\theta] = A\,\mathrm{var}[\theta]A^\top$

$p(\theta|X,y) = \mathcal{N}(m_N, S_N)$

$y_* = \phi^\top(x_*)\theta + \varepsilon$

$\mathbb{E}[y_*|x_*] = \mathbb{E}[\phi^\top(x_*)\theta + \varepsilon] = \underbrace{\phi^\top(x_*)}_{\text{const}}\underbrace{\mathbb{E}[\theta]}_{=m_N} + \underbrace{\mathbb{E}[\varepsilon]}_{=0} = \phi^\top(x_*)m_N$

$\mathrm{var}[y_*|x_*] = \mathrm{var}[\phi^\top(x_*)\theta + \varepsilon] = \mathrm{var}[\phi^\top(x_*)\theta] + \underbrace{\mathrm{var}[\varepsilon]}_{=\sigma^2}$

$= \phi^\top(x_*)\underbrace{\mathrm{var}[\theta]}_{S_N}\phi(x_*) + \sigma^2 = \phi^\top(x_*)S_N\phi(x_*) + \sigma^2$

Parameter Posterior and Predictions

- Prior $p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{m}_0, \boldsymbol{S}_0\right)$ is Gaussian ⏩ posterior is Gaussian:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) &= \mathcal{N}\left(\boldsymbol{m}_N, \boldsymbol{S}_N\right) \\
\boldsymbol{S}_N &= (\boldsymbol{S}_0^{-1} + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} \\
\boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y})
\end{aligned}
$$

- Mean $\boldsymbol{m}_N$ identical to MAP estimate

Parameter Posterior and Predictions

- Prior $p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{m}_0, \boldsymbol{S}_0\right)$ is Gaussian ⏩ posterior is Gaussian:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) &= \mathcal{N}\left(\boldsymbol{m}_N, \boldsymbol{S}_N\right) \\
\boldsymbol{S}_N &= (\boldsymbol{S}_0^{-1} + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} \\
\boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y})
\end{aligned}
$$

- Mean $\boldsymbol{m}_N$ identical to MAP estimate
- Assume a Gaussian distribution $p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{m}_N, \boldsymbol{S}_N\right)$. Then

$$
p(y_*|\boldsymbol{x}_*) = \mathcal{N}\left(y \,|\, \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{m}_N, \; \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x}_*) + \sigma^2\right)
$$

Parameter Posterior and Predictions

- Prior $p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{m}_0, \boldsymbol{S}_0\right)$ is Gaussian ⏩ posterior is Gaussian:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) &= \mathcal{N}\left(\boldsymbol{m}_N, \boldsymbol{S}_N\right) \\
\boldsymbol{S}_N &= (\boldsymbol{S}_0^{-1} + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} \\
\boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y})
\end{aligned}
$$

- Mean $\boldsymbol{m}_N$ identical to MAP estimate
- Assume a Gaussian distribution $p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{m}_N, \boldsymbol{S}_N\right)$. Then

$$
p(y_*|\boldsymbol{x}_*) = \mathcal{N}\left(y \,|\, \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{m}_N, \; \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x}_*) + \sigma^2\right)
$$

- $\boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x}_*)$: Accounts for parameter uncertainty in predictive variance

  More details ⏩ `https://mml-book.com`, Chapter 9

Marginal Likelihood

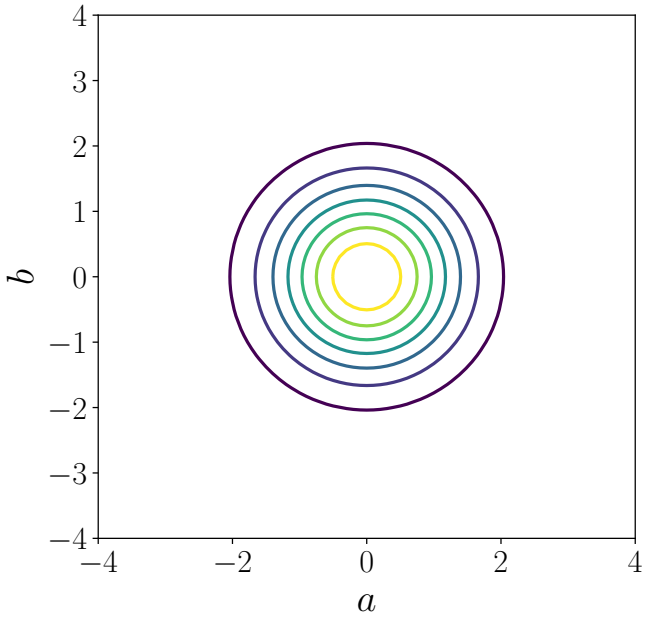
- Marginal likelihood can be computed analytically.
- With $p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \mathcal{N}\left(\boldsymbol{y} \,|\, \boldsymbol{\Phi}\boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top + \sigma^2 \boldsymbol{I}\right)$$

- Derivation via completing the squares (see Section 9.3.5 of MML book)

Consider a linear regression setting

$$y = f(x) + \epsilon = a + bx + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0, \sigma_n^2\right)$$
$$p(a, b) = \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\right)$$



Sampling from the Prior over Functions

Consider a linear regression setting

$$
\begin{aligned}
y &= f(x) + \epsilon = a + bx + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma_n^2\right) \\
p(a, b) &= \mathcal{N}\left(\mathbf{0},\, \boldsymbol{I}\right) \\
f_i(x) &= a_i + b_i x, \quad [a_i, b_i] \sim p(a, b)
\end{aligned}
$$

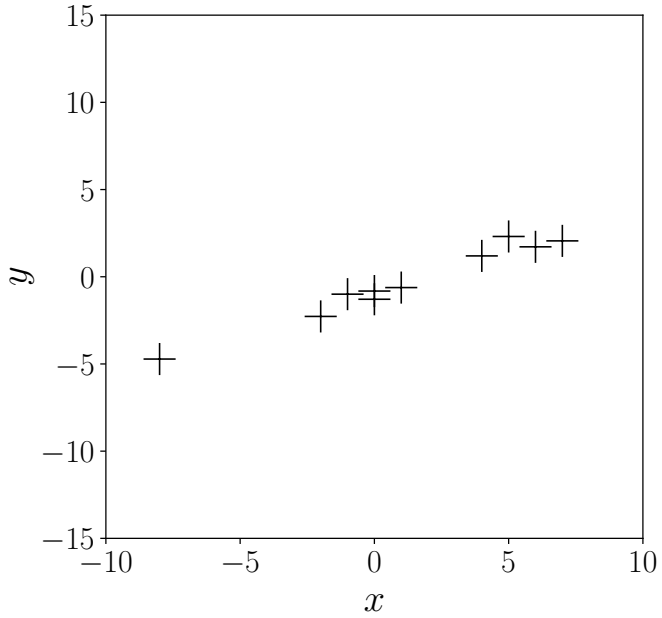
Sampling from the Posterior over Functions

Consider a linear regression setting

$$y = f(x) + \epsilon = a + bx + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma_n^2\right)$$

$$p(a, b) = \mathcal{N}\left(\mathbf{0},\, \boldsymbol{I}\right)$$

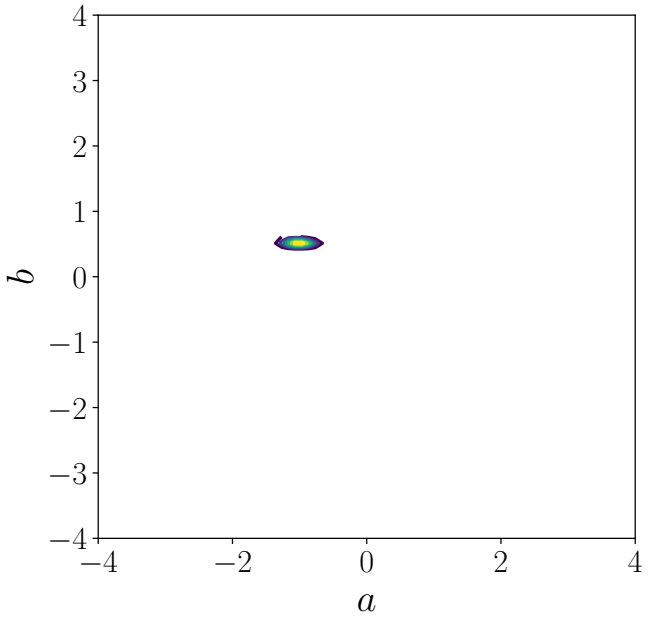
$$\boldsymbol{X} = [x_1, \ldots, x_N],\ \boldsymbol{y} = [y_1, \ldots, y_N]$$ Training inputs/targets



Sampling from the Posterior over Functions

Consider a linear regression setting

$$y = f(x) + \epsilon = a + bx + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma_n^2\right)$$
$$p(a, b) = \mathcal{N}\left(\mathbf{0},\, \boldsymbol{I}\right)$$
$$p(a, b|\boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{m}_N,\, \boldsymbol{S}_N\right) \qquad \text{Posterior}$$



Sampling from the Posterior over Functions

Consider a linear regression setting

$$
\begin{aligned}
y = f(x) + \epsilon = a + bx + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0, \sigma_n^2\right) \\
[a_i, b_i] \sim p(a, b | \boldsymbol{X}, \boldsymbol{y}) \\
f_i = a_i + b_i x
\end{aligned}
$$

Fitting Nonlinear Functions

- Fit nonlinear functions using (Bayesian) linear regression: Linear combination of nonlinear features

Fitting Nonlinear Functions

- Fit nonlinear functions using (Bayesian) linear regression: Linear combination of nonlinear features
- Example: Radial-basis-function (RBF) network

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \theta_i \phi_i(\boldsymbol{x}) \,, \quad \theta_i \sim \mathcal{N}\left(0, \sigma_p^2\right)$$

Fitting Nonlinear Functions

- Fit nonlinear functions using (Bayesian) linear regression: Linear combination of nonlinear features
- Example: Radial-basis-function (RBF) network

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \theta_i \phi_i(\boldsymbol{x}) \,, \quad \theta_i \sim \mathcal{N}\left(0, \sigma_p^2\right)$$

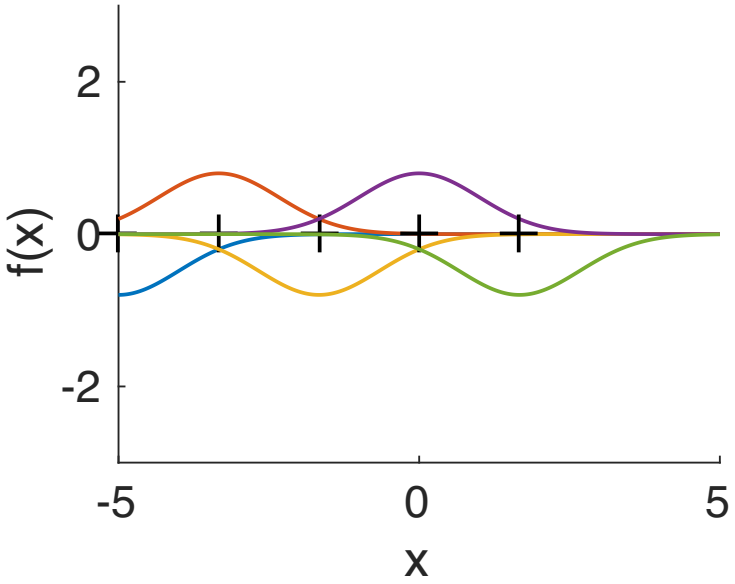
where

$$\phi_i(\boldsymbol{x}) = \exp\left(-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top (\boldsymbol{x} - \boldsymbol{\mu}_i)\right)$$

for given "centers" $\boldsymbol{\mu}_i$

Illustration: Fitting a Radial Basis Function Network

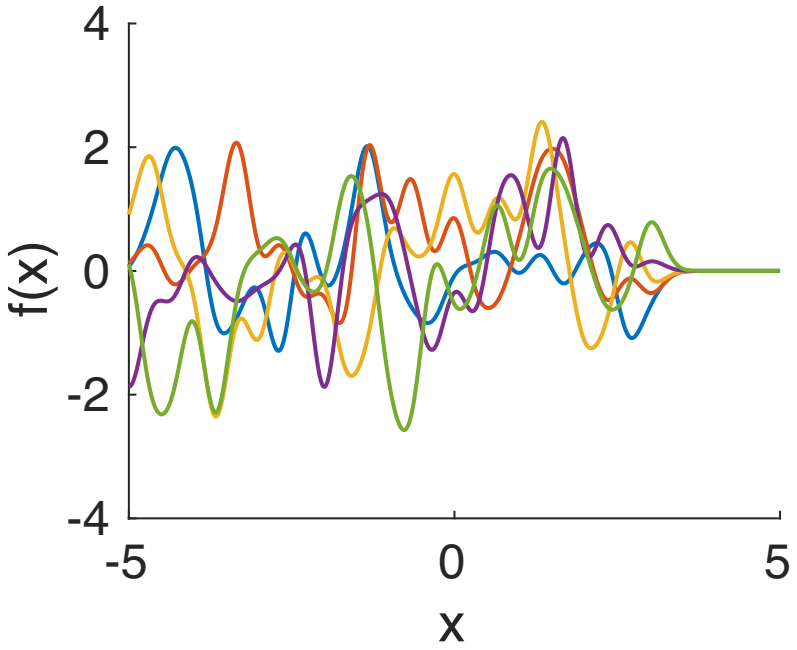
$$\phi_i(\boldsymbol{x}) = \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)$$



- Place Gaussian-shaped basis functions ϕ_i at 25 input locations μ_i, linearly spaced in the interval $[-5, 3]$

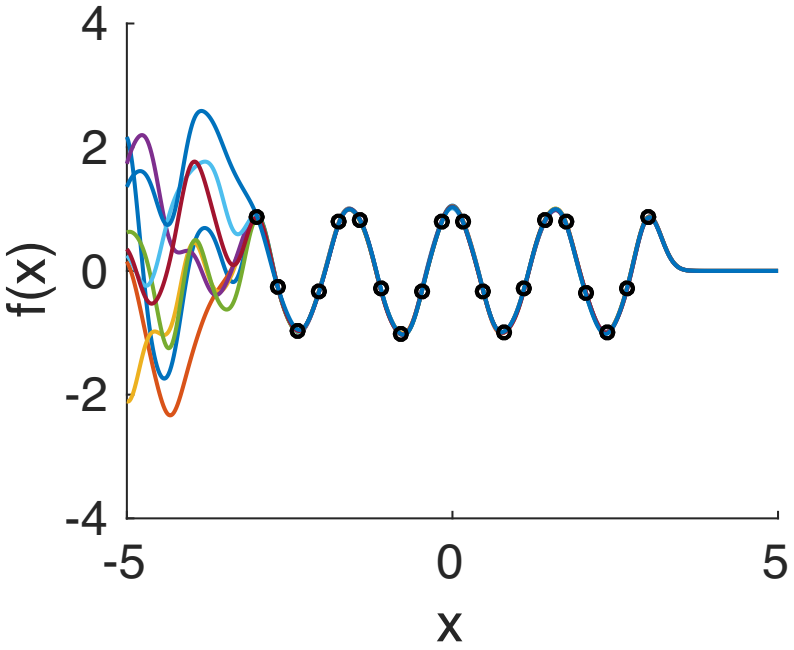
Samples from the RBF Prior

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \theta_i \phi_i(\boldsymbol{x}) \,, \quad p(\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\right)$$

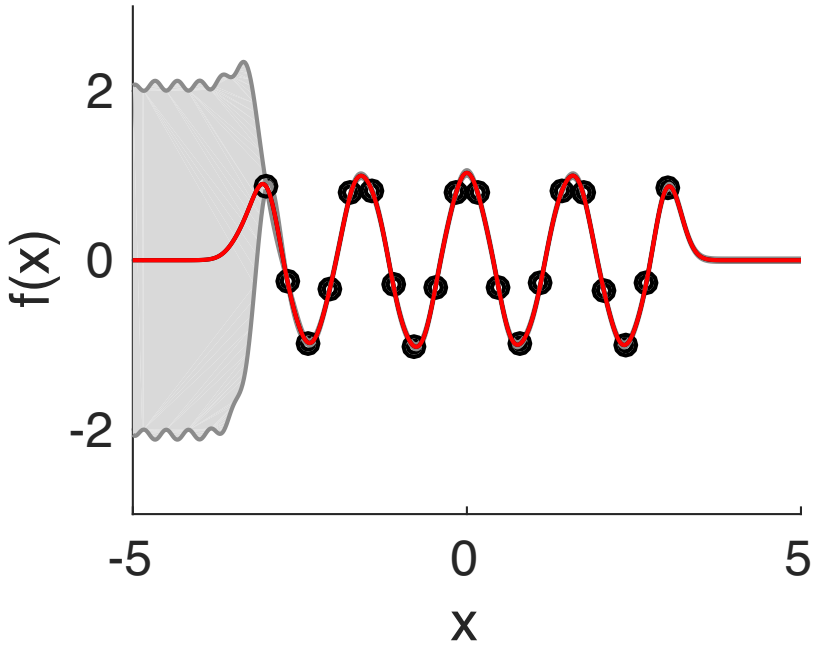


Samples from the RBF Posterior

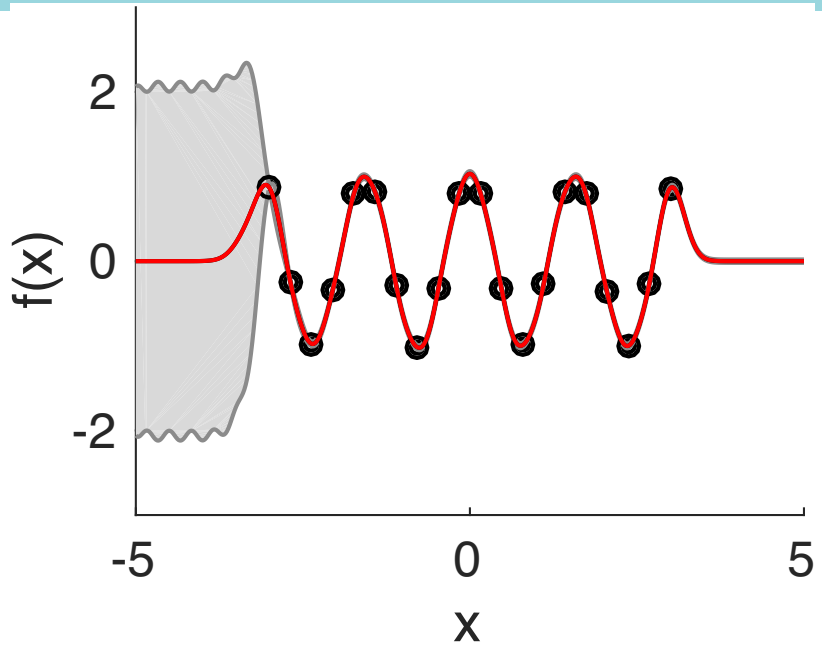
$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \theta_i \phi_i(\boldsymbol{x}), \quad p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{m}_N, \boldsymbol{S}_N\right)$$



RBF Posterior



Limitations



- Feature engineering (what basis functions to use?)
- Finite number of features:
  - Above: Without basis functions on the right, we cannot express any variability of the function
  - Ideally: Add more (infinitely many) basis functions

Approach

- Instead of sampling parameters, which induce a distribution over functions, sample functions directly
  - Place a prior on functions
  - Make assumptions on the distribution of functions

Approach

- Instead of sampling parameters, which induce a distribution over functions, sample functions directly
  - ⏩ Place a prior on functions
  - ⏩ Make assumptions on the distribution of functions
- Intuition: function = infinitely long vector of function values
  - ⏩ Make assumptions on the distribution of function values

Approach

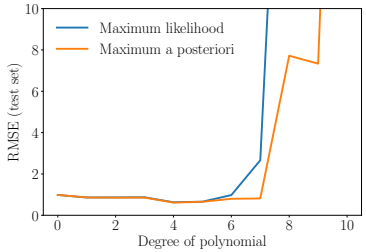
- Instead of sampling parameters, which induce a distribution over functions, sample functions directly
  - ⏩ Place a prior on functions
  - ⏩ Make assumptions on the distribution of functions
- Intuition: function = infinitely long vector of function values
  - ⏩ Make assumptions on the distribution of function values

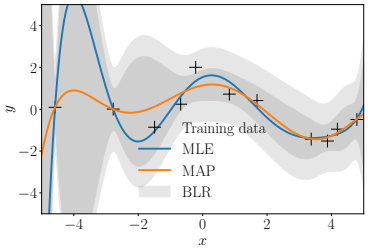
Approach

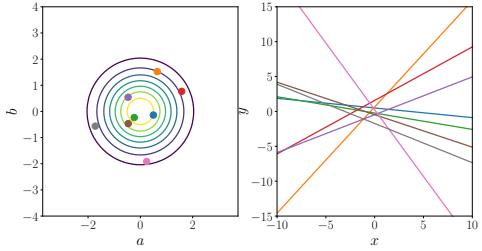
- Instead of sampling parameters, which induce a distribution over functions, sample functions directly
  - ⏩ Place a prior on functions
  - ⏩ Make assumptions on the distribution of functions
- Intuition: function = infinitely long vector of function values
  - ⏩ Make assumptions on the distribution of function values

⏩ **Gaussian process**

Summary







- Regression = curve fitting
- Linear regression = linear in the parameters
- Parameter estimation via maximum likelihood and MAP estimation can lead to overfitting
- Bayesian linear regression addresses this issue, but may not be analytically tractable
- Predictive uncertainty in Bayesian linear regression explicitly accounts for parameter uncertainty
- Distribution over parameters ⏩ Distribution over functions

Appendix

Joint Gaussian Distribution

- Joint Gaussian distribution

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right)$$

Joint Gaussian Distribution

- Joint Gaussian distribution

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right)$$

- Marginal:

$$\begin{aligned} p(\boldsymbol{x}) &= \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} \\ &= \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}) \end{aligned}$$

Joint Gaussian Distribution

- Joint Gaussian distribution

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right)$$

- Marginal:

$$\begin{aligned} p(\boldsymbol{x}) &= \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} \\ &= \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}) \end{aligned}$$

- Conditional:

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) &= \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \\ \boldsymbol{\mu}_{x|y} &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) \\ \boldsymbol{\Sigma}_{x|y} &= \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} \end{aligned}$$

Linear Transformation of Gaussian Random

If $\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ and $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ then

$$p(\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{z} \,|\, \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\right)$$

Product of Two Gaussians

$\boldsymbol{x} \in \mathbb{R}^D$. Then:

$$\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{a},\, \boldsymbol{A}\big)\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{b},\, \boldsymbol{B}\big) = Z\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{c},\, \boldsymbol{C}\big)$$

$$\boldsymbol{C} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1}$$

$$\boldsymbol{c} = \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b})$$

$$Z = (2\pi)^{-\frac{D}{2}}|\boldsymbol{A} + \boldsymbol{B}| \exp\left(-\tfrac{1}{2}(\boldsymbol{a} - \boldsymbol{b})^\top(\boldsymbol{A} + \boldsymbol{B})^{-1}(\boldsymbol{a} - \boldsymbol{b})\right)$$

Product of Two Gaussians

$\boldsymbol{x} \in \mathbb{R}^D$. Then:

$$\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{a},\, \boldsymbol{A}\big)\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{b},\, \boldsymbol{B}\big) = Z\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{c},\, \boldsymbol{C}\big)$$

$$\boldsymbol{C} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1}$$

$$\boldsymbol{c} = \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b})$$

$$Z = (2\pi)^{-\frac{D}{2}}|\boldsymbol{A} + \boldsymbol{B}| \exp\left(-\tfrac{1}{2}(\boldsymbol{a} - \boldsymbol{b})^\top (\boldsymbol{A} + \boldsymbol{B})^{-1}(\boldsymbol{a} - \boldsymbol{b})\right)$$

- Product of two Gaussians is an unnormalized Gaussian

Product of Two Gaussians

$\boldsymbol{x} \in \mathbb{R}^D$. Then:

$$\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{a},\, \boldsymbol{A}\big)\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{b},\, \boldsymbol{B}\big) = Z\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{c},\, \boldsymbol{C}\big)$$
$$\boldsymbol{C} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1}$$
$$\boldsymbol{c} = \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b})$$
$$Z = (2\pi)^{-\frac{D}{2}}|\boldsymbol{A} + \boldsymbol{B}| \exp\left(-\tfrac{1}{2}(\boldsymbol{a} - \boldsymbol{b})^\top(\boldsymbol{A} + \boldsymbol{B})^{-1}(\boldsymbol{a} - \boldsymbol{b})\right)$$

- Product of two Gaussians is an unnormalized Gaussian
- The "un-normalizer" Z has a Gaussian functional form:

$$Z = \mathcal{N}\big(\boldsymbol{a} \,|\, \boldsymbol{b},\, \boldsymbol{A} + \boldsymbol{B}\big) = \mathcal{N}\big(\boldsymbol{b} \,|\, \boldsymbol{a},\, \boldsymbol{A} + \boldsymbol{B}\big)$$

Note: This is not a distribution (no random variables)

Example: Marginalization of a Product

$$p_1(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{a},\, \boldsymbol{A}\right)$$
$$p_2(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{b},\, \boldsymbol{B}\right)$$

Then

$$\int p_1(\boldsymbol{x}) p_2(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \qquad \in \mathbb{R}$$

Note: In this context, $\mathcal{N}$ is used to describe the functional relationship between $\boldsymbol{a}, \boldsymbol{b}$. Do not treat $\boldsymbol{a}$ or $\boldsymbol{b}$ as random variables—they are both deterministic quantities.

Example: Marginalization of a Product

$$p_1(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{a},\, \boldsymbol{A}\right)$$
$$p_2(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{b},\, \boldsymbol{B}\right)$$

Then

$$\int p_1(\boldsymbol{x}) p_2(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = Z = \mathcal{N}\left(\boldsymbol{a} \,|\, \boldsymbol{b},\, \boldsymbol{A} + \boldsymbol{B}\right) \in \mathbb{R}$$

Note: In this context, $\mathcal{N}$ is used to describe the functional relationship between $\boldsymbol{a}, \boldsymbol{b}$. Do not treat $\boldsymbol{a}$ or $\boldsymbol{b}$ as random variables—they are both deterministic quantities.